THE ROYAL SOCIETY

Load minimization of the genetic code: history does not explain the pattern

Stephen J. Freeland[1*] **and Laurence D. Hurst**[2]

[1]*Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK*
[2]*Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK*

The average effect of errors acting on a genetic code (the change in amino-acid meaning resulting from point mutation and mistranslation) may be quantified as its 'load'. The natural genetic code shows a clear property of minimizing this load when compared against randomly generated variant codes. Two hypotheses may be considered to explain this property. First, it is possible that the natural code is the result of selection to minimize this load. Second, it is possible that the property is an historical artefact. It has previously been reported that amino acids that have been assigned to codons starting with the same base come from the same biosynthetic pathway. This probably reflects the manner in which the code evolved from a simpler code, and says more about the physicochemical mechanisms of code assembly than about selection. The apparent load minimization of the code may therefore follow as a consequence of the fact that the code could not have evolved any other way than to allow biochemically related amino acids to have related codons. Here then, we ask whether this 'historical' force alone can explain the efficiency of the natural code in minimizing the effects of error. We therefore compare the error-minimizing ability of the natural code with that of alternative codes which, rather than being a random selection, are restricted such that amino acids from the same biochemical pathway all share the same first base. We find that although on average the restricted set of codes show a slightly higher efficiency than random ones, the real code remains extremely efficient relative to this subset $p=0.0003$. This indicates that for the most part historical features do not explain the load-minimization property of the natural code. The importance of selection is further supported by the finding that the natural code's efficiency improves relative to that of historically related codes after allowance is made for realistic mutational and mistranslational biases. Once mistranslational biases have been considered, fewer than four per 100 000 alternative codes are better than the natural code.

Keywords: genetic code; evolution; natural selection; error minimization; historical restriction

1. INTRODUCTION

The universal genetic code is structured such that when errors occur (such as mistranslation or mutation), the resulting change in meaning is to an amino acid of very similar properties to the one that should have been there, at least in terms of polarity (Epstein 1966; Fitch 1966; Goldberg & Wittes 1966; Woese *et al*. 1966; Alff-Steinberger 1969; DiGiulio 1989*a*,*b*; Haig & Hurst 1991; Szathmary & Zintzaras 1992; Ardell 1998; Freeland & Hurst 1998; see DiGiulio (1997) for a review). Quantification of this property (termed the 'load' of the code; see Szathmary & Maynard Smith 1995) shows that the overwhelming majority of possible codes are less efficient in this respect such that the probability of arriving at a genetic code that is as or more efficient than the natural code by chance alone is of the order $p=0.0001$ (Haig & Hurst 1991; Ardell 1998; Freeland & Hurst 1998): the load of the code is significantly lower than can be explained by chance.

Two classes of explanation may be provided to account for this property (Crick 1968; and see DiGiulio (1997) for a description of the development of these hypotheses). First, it might be that selection between alternative codes has resulted in the near universality of one that is extremely efficient at error minimization (e.g. Sonneborn 1965; Woese 1965; Fitch 1966; Fitch & Upper 1987; DiGiulio 1989*a*,*b*; others are described in DiGiulio (1997)). That selection might well underlie the pattern is suggested by the fact that the perceived efficiency of the natural code increases when the method of quantification is adjusted to include recognized biases in both mutation and mistranslation (Freeland & Hurst 1998). A selective argument is also supported by the finding that (as required) variation in codes is possible and observed (see, for example, Jukes & Osawa 1991). The mechanism of codon reassignments has attracted considerable attention (see, for example, Osawa & Jukes 1989, 1995; Osawa *et al*. 1992; Schultz & Yarus 1994, 1996; Szathmary & Maynard Smith 1995; Jukes & Osawa 1997; Santos *et al*. 1997; Yarus & Schultz 1997).

An alternative hypothesis is that the apparent load minimization may be largely an artefact, produced by what might be considered 'historical' forces (Nirenberg *et al*. 1963; Pelc & Welton 1966; Dillon 1973; Wong 1975, 1976, 1980, 1981, 1988; Wong & Bronskill 1979; Taylor &

*Author for correspondence (s.freeland@gen.cam.ac.uk).

Received 27 May 1998 *Accepted* 20 July 1998

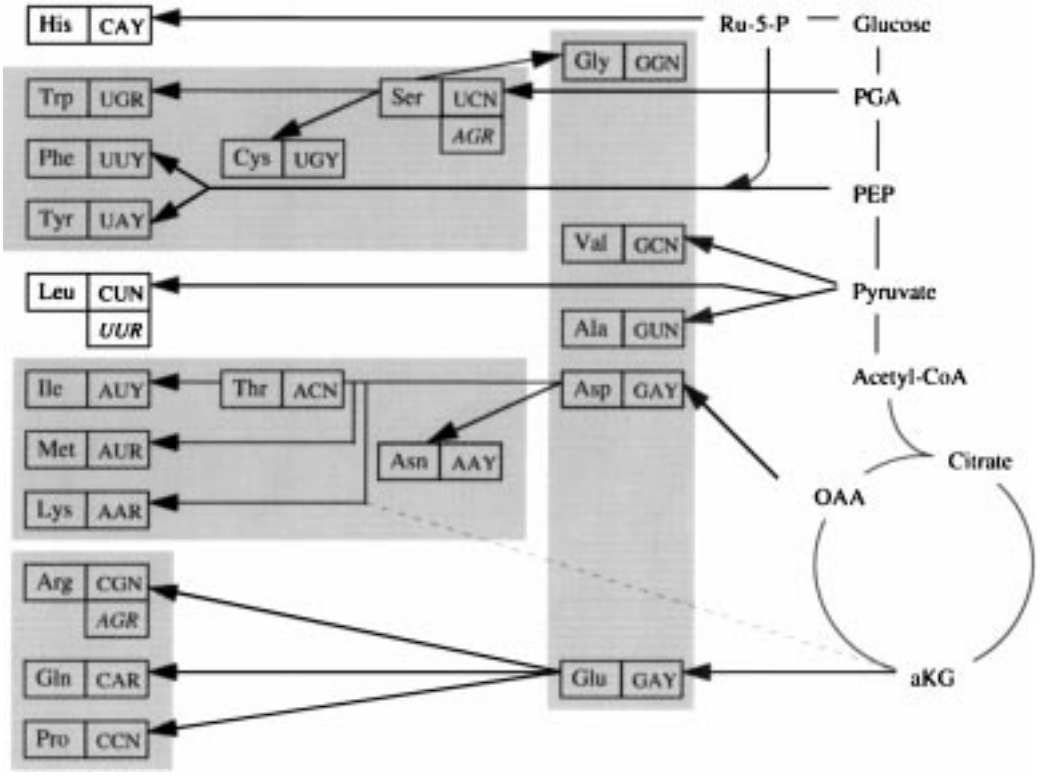

Figure 1. Biosynthetic pathways and codon assignments for the 20 amino acids coded for by the canonical genetic code, adapted from Taylor & Coates (1989). Grey-shaded areas highlight codon assignments that suggest a relationship between codon families and amino-acid meanings. Consistent with the majority of associated literature, in addition to the four specific nucleotide abbreviations (U, C, A and G) we use N to refer to 'any nucleotide', Y to refer to 'any pyrimidine' (U or C) and R to refer to 'any purine' (A or G).

Coates 1989). Typically, such arguments propose that the primordial genetic code contained fewer than the 20 amino acids seen today (i.e. that the primordial code contained a higher level of redundancy). Subsequent evolution thus merely split existing synonymous blocks into subsets coding for both the original amino acid and a biosynthetically related variant (e.g. Hartman 1975; Wong 1975, 1980; Wong & Bronskill 1979; Jukes 1983; Szathmary & Zintzaras 1992; Szathmary 1993; Bashford *et al*. 1998). If biosynthetically related amino acids also share physiochemical properties, such that similar codons code for similar amino acids for historical reasons, then any quantification of the relationship of codon assignments is likely to reflect this (e.g. DiGiulio 1996). Under this hypothesis, error minimization through natural selection is limited in scope within the framework of deterministic biosynthetic links between codons and amino acids (Wong 1980).

One potential weakness of these 'historical' arguments is that perceived patterns of biosynthetic relatedness between the amino acids assigned to similar codons may themselves be artefacts, resulting from the fact that most amino acids are biosynthetically related. Specifically, apparent patterns of 'precursor-product' amino-acid pairings are found to occur when the structure of the code is permuted randomly (see Amirnovin (1997) for a direct comment on Wong's (1980) approach). Nonetheless, one of the clearest and most general observations of biosynthetic relatedness is that amino acids which share a biosynthetic pathway tend to also share the same nucleotide identity in the first base position of their corresponding codons (Taylor & Coates 1989). More specifically, the 'shikimate' (aromatic) family of amino acids (Trp, Phe and Tyr), together with their biosynthetic precursors Ser and Cys, are coded by UNN, the 'glutamate' family (Gln, Pro and Arg) are coded by CNN. The 'aspartate' family (Ile, Met, Thr, Lys and Asn) are coded for by ANN. Codons starting with a guanine (i.e. GNN) are from different families (Glu, Asp, Ala, Val, Gly), but all appear at or near the head of a biosynthetic pathway (figure 1).

Such patterns are good evidence that when new amino acids were incorporated into the code, the codons they trapped were more likely to be those of biosynthetically related amino acids. Can such a historical explanation account for the error-minimizing property of the genetic code? A consideration of this pattern suggests that the historical effect does not necessarily offer an alternative to the adaptive hypothesis. The general pattern described by Taylor & Coates (1989) does not explain the codon assignments of histidine and leucine, nor does it explain the assignments for the paired codons additional to the 'family box' assignments for those amino acids that possess six synonyms (leucine, arginine and serine). More fundamentally, it does not explain specific codon assignments within these subgroupings. That codons can be reassigned also suggests that historical forces need not determine all properties of the code. Conversely, it is entirely plausible that the set of possible codes defined by this historical restriction shows a generally higher level of load minimization than the much larger, unrestricted set of possible codes (see § 3): indeed, the disparity in size of the two sets is such that they may be regarded as effectively independent.

The issue, therefore, is by no means resolved, and it is worthwhile to ask the extent to which proposed historical forces explain the previously reported efficiency of the natural code. The previous method of measuring the relative efficiency of the code compared it against a sample of random variants in which the amino-acid

assignments of synonymous codon sets were allowed to vary freely (Haig & Hurst 1991; Ardell 1998; Freeland & Hurst 1998). Here we ask how the natural code compares with a set of alternative codes comprising only 'historically reasonable' permutations. We additionally compare the results from our previous, unconstrained code analysis (Freeland & Hurst 1998) with results from the historically restricted set.

In constructing our 'historically reasonable' set of possible codes we follow Taylor & Coates's (1989) finding, i.e. we restrict the world of possible codes to those in which codon block assignments are allowed to vary only within the biosynthetic pathways described above. For example, variant codes are produced in which the synonymous codon sets starting with a U are only allowed to take a shikimate assignment (Ser, Cys, Trp, Phe or Tyr). In our previous analysis codons starting with U were unrestricted as to which amino acids they might code for.

If the load-minimization property of the natural code were solely a product of the way it was assembled, then a significant proportion of all historically restricted codes should show a similar level of efficiency. Were the natural code's load-minimization property the result of selection, then the natural code should remain special within the subset of historically restricted codes.

2. METHODS

Our method of quantifying the efficiency of a given genetic code is the 'mean square' (MS) measure used in previous models (DiGiulio 1989*b*; Haig & Hurst 1991; Ardell 1998; Freeland & Hurst 1998). This measure calculates the mean squared difference in amino-acid property resulting from all possible single-base errors to all codons within a code: synonymous changes are included in the calculation (see DiGiulio (1989*b*) and Ardell (1998) for a discussion of this point), but changes to and from termination codons are ignored. The overall MS measure of a code (the MS0 value) is partitioned into three component scores, representing the mean squared change resulting from all errors in each of the codon base positions (thus MS1, MS2 and MS3 refer to the average effect of errors occurring at the first, second and third codon bases, respectively).

All MS calculations presented here consider a single attribute of amino acids, 'polar requirement' (Woese *et al*. 1966), an empirically determined measure of amino-acid similarity based on polarity. This we employ, as earlier analyses have identified its probable evolutionary significance in this context (e.g. DiGiulio 1989*a*; Haig & Hurst 1991; Szathmary & Zintzaras 1992). Indeed, under a crude quantification of mistranslation biases, this parameter suggests that the probability of evolving a 'better' code is one in a million (Freeland & Hurst 1998). It is so unlikely that such a figure would appear artefactually that we assume that this parameter is informative of something.

In addition to the calculation of the basic MS measures, we calculate an arithmetic weighted mean square (wMS) measure which incorporates a weighting variable representing the relative importance (i.e. relative frequency) of transition and transversion errors when quantifying the average effect of changing a codon base (Freeland & Hurst 1998). In calculating the MS value of each codon, the weighting variable is only applied to changes in codon meanings resulting from transition errors (i.e. differences in codon meaning resulting from a purine-to-purine error or pyrimidine-to-pyrimidine error); transversion errors are always given a weighting of one. Thus a wMS measure at a weighting of two is calculated such that squared individual differences resulting from transition mistakes are being weighted twice as heavily as squared differences resulting from transversion errors. It should thus be noted that wMS measures at a weighting of one are exactly the same as basic MS measures.

The random sample of variant genetic codes against which the canonical code was tested was generated according to the following rules.

1. All random variants maintain the same synonymous block structure as the canonical code (for example, the block UUN is always subdivided into two synonymous blocks: UUY is assigned one amino acid, and UUR is assigned another). Maintaining a constant superstructure for all variant codes controls for the level of redundancy inherent in the canonical code: this is important because the MS measure of code efficiency includes silent mutations and is thus directly affected by the amount and pattern of redundancy.
2. The 20 'sense' synonymous blocks which together comprise the genetic code are divided into four groups, each group comprising the synonymous codon sets described by Taylor & Coates (1989) which share a common base identity at the first codon position (UNN, CNN, ANN and GNN). In the cases where synonymous codon sets comprise six codons, it is assumed that the amino-acid assignment of the extra pair of codons is dictated by the assignment of the four-codon family box (for example, the meaning of the codon pair AGY is always determined by the assignment of codon UCN). The codon assignments of synonymous blocks are allowed to vary within groups, but not between groups. Thus, for example, the five synonymous blocks contained within UNN codons always code for the five shikimate amino acids, and the extra pair AGY is thus also always assigned the shikimate amino acid coded for by codon set UCN. The extra pair UUR is, however, assigned the same meaning as codon block CUN.

These rules are illustrated in figure 2, and were incorporated into an ANSI 'C' program, which generated a random sample of one million variant codes and measured the wMS values of each over a weighting range of 1 (transitions and transversions weighted equally) to 20 (transitions weighted 20 times more heavily than transversions).

3. RESULTS

(a) *Independence from previous studies*

The set of possible genetic codes analysed in previous studies of this type (Haig & Hurst 1991; Freeland & Hurst 1998) maintains the redundancy structure of the natural genetic code but allows amino-acid assignments to vary freely. This unrestricted set thus comprises $20! = 2\,432\,902\,008\,176\,640\,000$ possible permutations of the natural genetic code. The set of possible codes analysed here (the restricted set) comprises $(5!)^4 = 207\,360\,000$ possible permutations (figure 2). All members of the restricted set are also members of the unrestricted set. The highest recorded estimate of optimization of the natural code (Freeland & Hurst 1998), based on the unrestricted set, is that one in a million randomly chosen codes has a lower MS0 value. Thus, even if this remarkable result were accepted at face value, we would

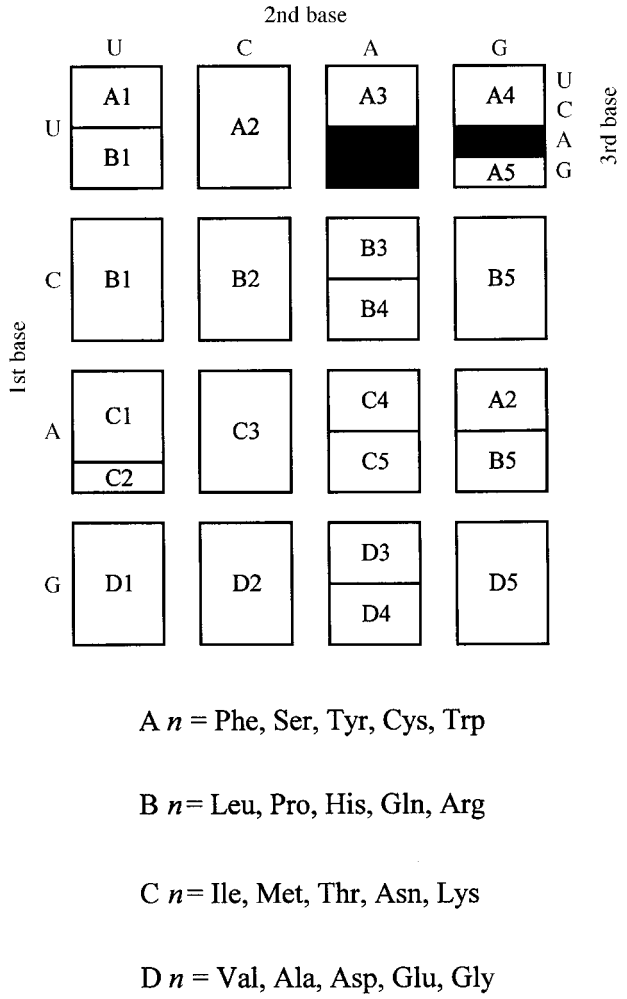

Figure 2. Rules for generating variant genetic codes consistent with the Taylor–Coates model of historical constraint. Codon assignments are divided into four groups (A to D), each containing five members (1 to 5). Codon assignments were allowed to vary randomly within members of a group but not between groups. Thus, for example, amino acids Phe, Ser, Tyr, Cys and Trp are randomly assigned one synonymous block each of elements A1 to A5.

predict the existence of (20!)/1 000000 = 2 432 902 008 177 better codes within the unrestricted set. The subset of better codes within the unrestricted set could contain the entire set of historically restricted codes several thousand times over: it would be logically possible that every single 'historically restricted' code had a lower MS0 value than the natural genetic code.

Put another way, the expected proportion of 'historically restricted' codes generated by chance alone in the previously analysed sample of one million unrestricted codes (Freeland & Hurst 1998) is 0.000085. In other words, the sample produced here under a model of historical restriction is independent from the sample analysed previously.

(b) *Equal transition/transversion bias*

Table 1 compares the MS values calculated for the sample of one million variant codes generated under the historically restricted model with the corresponding values produced under the previous (unrestricted) model (Freeland & Hurst 1998). At each individual codon position (MS1, MS2 and MS3), and at all codon positions combined (MS0), the sample of variant codes produced under the restricted model shows less variation than that produced under the unrestricted model, as perhaps might be expected. Furthermore, at the second and third codon positions (MS2 and MS3) and at all codon positions combined (MS0), the restricted sample possesses lower mean MS values than the unrestricted sample, whereas at the first codon position (MS1) the restricted sample shows a higher mean MS value than the unrestricted sample.

These observations are, to some extent, consistent with the idea that the MS measurement system reflects the finding that similar amino acids (defined as amino acids which share a biosynthetic pathway) are assigned similar codons (see also DiGiulio 1996). It should be noted, however, that the differences are all relatively small, suggesting that the effect of the historical forces on the perceived efficiency of the canonical code is rather weak.

A more direct assessment of the effect of historical forces on the perceived efficiency of the code comes from comparing the number of 'better' (lower MS) variant codes found within the two samples. Out of one million variant codes generated under the previous (unrestricted) model, only 114 were more conservative (i.e. gave a lower MS0) than the natural code, giving an estimated probability of arriving at a code as conservative as the canonical code by chance alone of $p = 0.0001$. Of the one million codes generated under the restricted model, 284 'better' (lower MS0) codes were found. Thus, when the historical restriction is incorporated into the model, the canonical genetic code still exhibits statistically highly significant evidence of adaptation for error minimization. Even if variation of codon assignments is limited to amino acids within a particular biochemical pathway, the chance of arriving at a genetic code as or more efficient than the canonical genetic code is estimated at $p = 0.0003$. Whether the difference between $p = 0.0001$ and $p = 0.0003$ has any biological meaning is unclear (see § 4).

It may also be noted that under the restricted model, the perceived relative efficiency of the first codon position appears five times better (more efficient) than under the unrestricted model, whereas the third codon position appears six times less efficient. The net result is that under the historical model, the perceived efficiency of the first and third codon positions is very similar, whereas under the unrestricted model the first codon position appears 30 times less efficient than the third codon position. Under both restricted and unrestricted models, the second base codon position superficially shows no evidence of adaptation for error minimization (being three orders of magnitude less efficient than first or third codon positions).

Finally, it may be noted that under the restricted model, the relative efficiency of all codon positions combined is lower than that for any individual base position. Upon further examination, this last result was revealed to come from the fact that many variant codes which achieve particularly good efficiency at any one codon position do so at the expense of efficiency at other codon positions.

(i) *Biases in the error process*

If the code were the result of selection we might expect that the incorporation of realistic biases in error processes (e.g. mistranslation errors, point-mutation errors) should

Table 1. *Descriptive statistics for the distributions of MS values formed by a sample of one million variant codes under constrained and unconstrained models*

measure		unconstrained model (n=1 000 000)	constrained model (n=1 000 000)	t' (p)[a]
mean (s.d.)	MS0	9.41 (1.51)	8.87 (0.95)	3029 ($\ll$0.01)
	MS1	12.04 (2.80)	12.38 (2.49)	90.81 ($\ll$0.01)
	MS2	12.63 (2.60)	11.24 (0.99)	499.0 ($\ll$0.01)
	MS3	3.59 (1.50)	3.02 (1.18)	298.9 ($\ll$0.01)
proportion of better codes found	MS0	0.0001	0.0003	—
	MS1	0.0030	0.0006	—
	MS2	0.2216	0.2466	—
	MS3	0.0001	0.0006	—

[a]The probabilities that the two samples are drawn from the same population (i.e. that the biosynthetic restriction rules make no difference to the mean MS value for random codes), were calculated using the t'-test (for difference of means with unequal sample variance) described in Sokal & Rohlf (1981).

reveal the code to be relatively even more efficient (Freeland & Hurst 1998). We have analysed the effects of incorporating both of these biases.

(c) *Incorporating a range of transition/transversion biases*

As with previous results for the unrestricted model, we further tested the MS values of the natural genetic code against those of a random sample of one million variants over a range of transition/transversion biases. Figure 3 compares the results under the two models. In each plot, the y-axis represents the proportion of random variant codes found which were more conservative (lower wMS values) than the actual genetic code, and the x-axis represents different weightings of transition/transversion bias. The first plot shows the proportion of more conservative codes at the second base position (wMS2) and at all base positions combined (wMS0). The second plot shows the same information for the first and third base positions (wMSl and wMS3) and for all base positions combined (wMS0).

Clear similarities exist between the results of the restricted and unrestricted models. Under both models, the relative efficiency of the natural code increases with a mild transition/transversion bias at all individual codon positions and at all codon positions combined. This is consistent with adaptive expectations. Furthermore, under the historical restriction, the strongest effect of increasing transition/transversion bias is once again seen in the second base position, where the relative efficiency of the natural code improves approximately sixfold when a transition bias of three is applied (i.e. when transitions are weighted three times more heavily than transversions).

One clear difference between the results of the two models is that under the restricted model, the second base position shows a clear peak in efficiency at a transition bias of three (worsening at higher transition biases), whereas in the unrestricted model, the efficiency of the second base position shows a consistent (though apparently asymptotic) increase in efficiency as the transition bias increases to infinity (figure 3*a*). Additionally, under the restricted model the perceived efficiency of the second base position is actually better than under the unrestricted model at mild transition weightings (between two and five). The biological significance of these two findings is unclear (see § 4).

Equally noteworthy is that the apparent lower overall relative efficiency of the natural code under the restricted model (measured in terms of the number of variants with a lower wMS0 value; see table 1) disappears when a transition weighting is applied. Indeed, the overall efficiency of the code is actually better under the restricted model than under the unrestricted model at transition weightings above two. Also, in contrast to the previous results, overall code relative efficiency under the restricted model remains remarkably consistent over all transition weightings above one.

Could the response of the natural code to transition/transversion biases be an artefact? Were this so, we would expect many of the codes that are better under no mutational bias to be more efficient under mutational bias. A comparison of the canonical genetic code with a subset of the superficially 'better' (lower MS0) variant codes indicates, as under the unrestricted model, that this is not so. The behaviour of the code under mutational biases is hence unlikely to be an artefact of genetic code superstructure (figure 4).

(d) *Mapping translational error data to MS calculations*

One of the previous adaptive studies extended the MS measurement system to represent a simple mapping of mistranslation parameters (Freeland & Hurst 1998). Under this model, the system used to calculate the MS value of each codon weighted the three individual codon base positions separately to reflect a suggested pattern of mistranslational bias (both in terms of different overall relative weightings for each base position and in terms of different transition/transversion weightings for each codon position). The quantification used to produce this 'translation MS' ('tMS') value is described in Freeland & Hurst (1998). We therefore mapped this weighting system into our model and re-tested the canonical code against a sample of one million random variants. The distribution of results obtained in this manner, together with the relative position of the natural code is shown in figure 5; a comparison of the results produced under restricted and unrestricted models is shown in table 2.

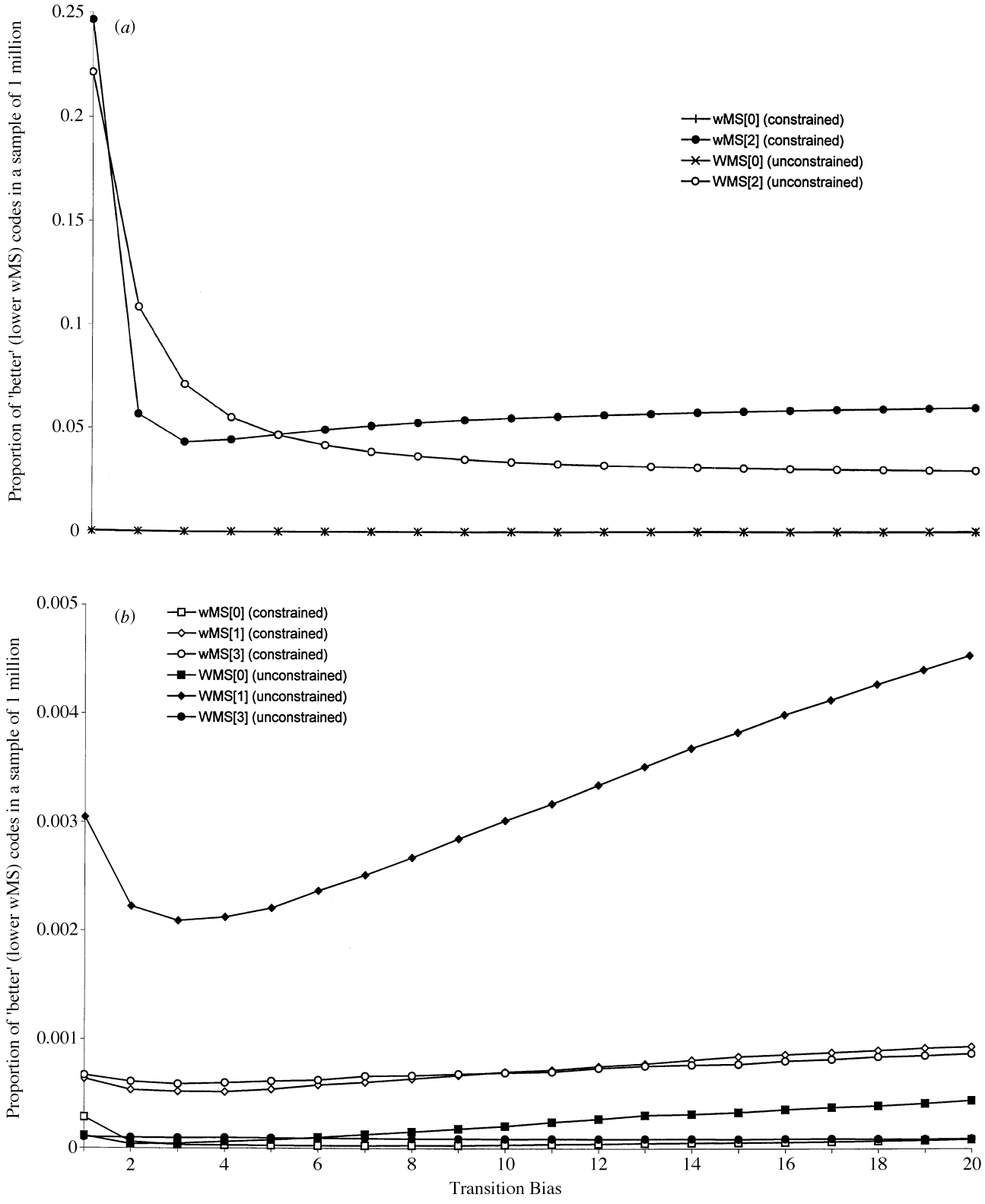

Figure 3. The proportion of 'better' (lower wMS) variant codes found (in a sample of one million) over a range of transition/transversion biases: comparison of constrained and unconstrained models. (*a*) Second base position (wMS2) and all bases combined (wMS0). (*b*) First (wMS1), third (wMS3) and all base positions combined (wMS0).

Under translational bias we find only 36 in a million codes to be more efficient that than the natural code. The difference between this level of efficiency and the level apparent when no bias is applied is comparable to the result previously reported when codon assignments are unrestricted (Freeland & Hurst 1998). This again strongly suggests that historical forces do not account for the natural code's ability to minimize the effects of mistranslation. However, previously only one code in a million was found to be better. Once again, the biological significance of these discrepancies is uncertain (see § 4).

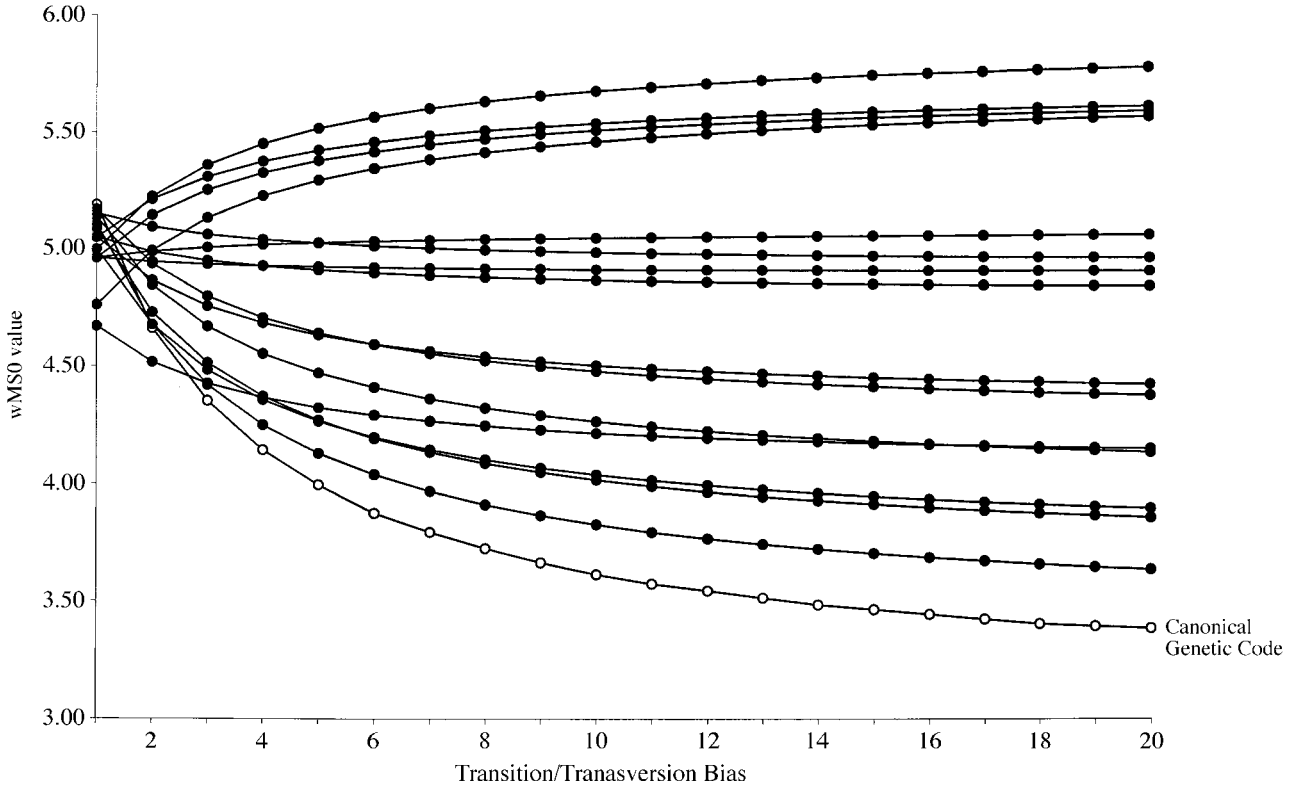

Figure 4. A comparison of wMS0 values for the natural genetic code and 15 superficially 'better' (lower MS0) variant codes under a range of transition/transversion biases.

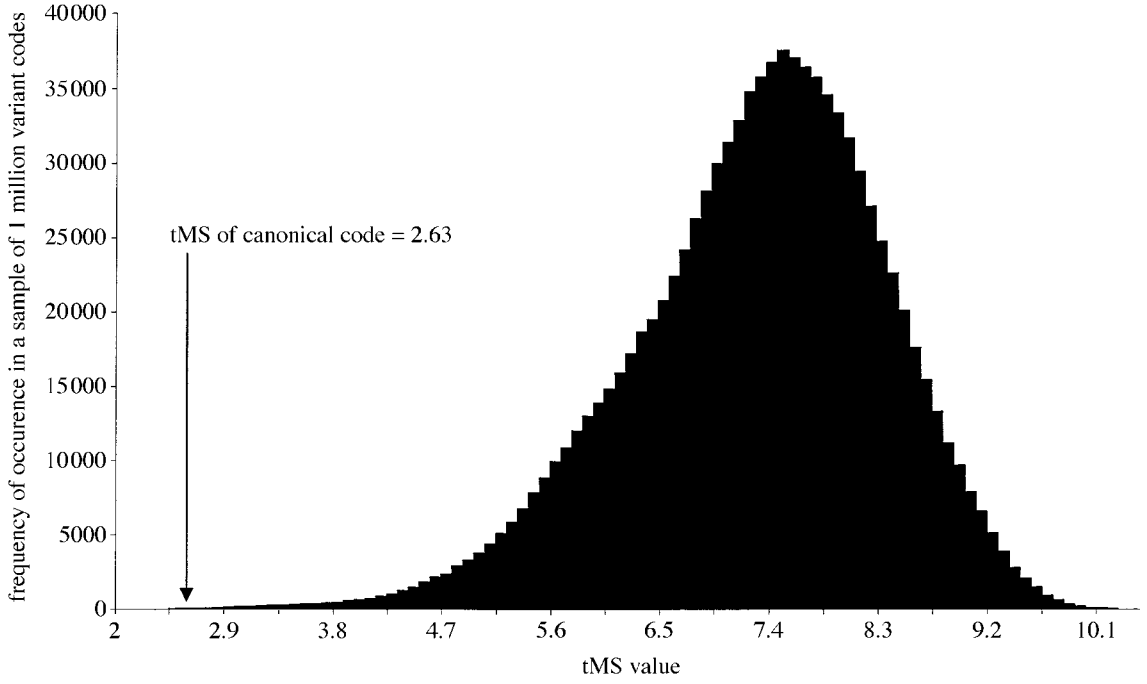

Figure 5. Frequency distribution for the tMS (eqivalent to MS0 adjusted for mistranslation parameters) values obtained from one million random variants generated under the restricted model of code variation. The *x*-axis gives a particular range of categories of MS values, and the *y*-axis gives the number of random variant codes generated with an MS value in that category. An arrow indicates the category into which the tMS value for the canonical code falls: the cumulative frequency to the left of this arrow therefore indicates the proportion of more conservative codes found among the random variants. This cumulative frequency is 37, indicating that under our quantification of mistranslation parameters, the probability of a code as efficient or more efficient than the natural code evolving by chance alone is $p=0.000037$.

Table 2. *Values of tMS calculated for the natural code, and for a sample of one million random variants, under the restricted and unrestricted models*

	canonical code	sample of one million random variants		
	tMS value	mean	s.d.	no.'better' codes found
constrained model	2.63	7.31	1.04	36
unconstrained model	2.63	7.63	1.35	1

4. DISCUSSION

Can history explain the natural genetic code's ability to minimize the effects of errors? Codons sharing the same first base identity do indeed tend to share biosynthetic pathways, and so such an historical effect might be likely. We have incorporated this restriction into the model by which plausible alternative genetic codes are generated, and demonstrated the independence of the set of possible genetic codes thus defined from previous adaptive studies. Such historically related codes do indeed appear to have a slightly lower mean absolute efficiency than do a comparable set of purely random codes, indicating a possible small effect of history in determining the code's ability. However, our previous results (Freeland & Hurst 1998) remain qualitatively unchanged. Within the set of historically related codes, the natural code may be regarded as highly adapted: the refined estimate for the probability of arriving at a code as or more efficient is between $p=0.0003$ and $p=0.00004$ depending on assumptions about biases in the error process. That the efficiency of the universal code improves relative to that of historically related codes after incorporation of realistic mutational and mistranslational biases into the calculation also strongly suggests a role for selection.

Although the above findings are distinct from those from previous unrestricted analyses, this result is not wholly unexpected. Most notably, previous analyses (Haig & Hurst 1991; Freeland & Hurst 1998) have shown that changes at the first site are relatively conservative, whereas those at second sites are not. Were the code's minimization property to result from the fact that related amino acids share the same first base, the opposite result would have been expected.

The analysis presented here is by no means the last word on this issue. As we have mentioned, it is remarkable that one parameter (polar requirement) seems to provide such highly significant statistics. It is hard to believe that differences in this one parameter explain all variation in protein (and hence code) fitness; there remains a possibility that the perceived 'adaptation' reflects (at least in part) deterministic stereochemical interactions between amino acids and their corresponding anticodons (e.g. DiGiulio 1996). Further, we use a mean-square method to assign fitness. We could have taken a mean modular difference or any modular power relationship of the difference (e.g. DiGiulio 1989*a*; Ardell 1998). We do not, however, know how difference in chemical property and fitness covary and hence do not know which measure is optimal. Also, our quantification of mistranslation bias is crude and should only be taken as a rough guide to biological significance.

For these reasons it is uncertain whether relatively small quantitative differences between the results of the comparison of the real code with restricted and unrestricted codes have any biological relevance. For both the analysis of relative efficiency with no biases and that with translational biases, the code appears marginally less significantly adapted when compared against the historically restricted set than when compared with the unrestricted set. In contrast, when measurements of code efficiency are transformed such that changes in codon meaning resulting from transition errors are weighted two or more times more heavily than those resulting from transversion errors, the code actually appears more highly adapted under the restricted model than under the unrestricted model (the perceived relative efficiency of the code improving an order of magnitude over this interval). However, the potential inaccuracy of our measurement system (based on squared differences of a single parameter value) urges caution. Is three per 10 000 importantly different from one per 10 000? Is 36 per million importantly different from one per million? The latter figure, being an order of magnitude difference, is the most suggestive of some biological relevance.

Given the limitations of the analysis, it is unwise to read too much into the data beyond the observation that the natural code compares against a historically restricted set of possible codes much as it does against an unrestricted set. For this reason we conclude that, although historical forces may well determine that amino acids in the same biosynthetic pathway share the same first codon, this does not explain the error-minimizing property of the natural code: a role for both history and selection is necessary (see also DiGiulio 1998).

The apparent lack of interference between the historical forces and selection is noteworthy. One reason for this might be that the amino acids that were actually incorporated into the code were those which were both biosynthetically related and allowed load minimization (Fitch & Upper 1987; Szathmary & Zintzaras 1992). We might conjecture that other possible amino acids produced by different biosynthetic pathways were not incorporated because they upset the pattern of error minimization. Such a model additionally has the possibility of explaining why only 20 amino acids are used. As a code increases its information content, error-minimization potential is likely to be reduced (Szathmary 1991, 1992; Szathmary & Maynard Smith 1995). Whether the natural set of 20 amino acids represents an optimal balance between enzymatic efficiency and susceptibility to error we leave to future investigation.

We thank John Barrett for helpful discussion of aspects of this work. The manuscript was improved by comments from Massimo DiGiulio and two anonymous referees.

REFERENCES

Alff-Steinberger, C. 1969 The genetic code and error transmission. *Proc. Natn. Acad. Sci. USA* **64**, 584–591.

Amirnovin, R. 1997 An analysis of the metabolic theory of the origin of the genetic code. *J. Molec. Evol.* **44**, 473–476.

Ardell, D. H. 1998 On error minimization in a sequential origin of the standard genetic code. *J. Molec. Evol.* **47**, 1–13.

Bashford, I., Tsohantjis, I. & Jarvis, P. D. 1998 A supersymmetric model for the evolution of the genetic code. *Proc. Natn. Acad. Sci. USA* **95**, 987–992.

Crick, F. H. C. 1968 The origin of the genetic code. *J. Molec. Biol.* **1968**, 367–379.

DiGiulio, M. 1989*a* Some aspects of the organisation of the genetic code. *J. Molec. Evol.* **29**, 191–201.

DiGiulio, M. 1989*b* The extension reached by the minimisation of the polarity distances during the evolution of the genetic code. *J. Molec. Evol.* **29**, 288–293.

DiGiulio, M. 1996 The beta sheets of proteins, the biosynthetic relationships between amino acids and the origin of the genetic code. *Origins Life Evol. Biosph.* **26**, 589–609.

DiGiulio, M. 1997 On the origin of the genetic code. *J. Theor. Biol.* **187**, 573–581.

DiGiulio, M. 1998 Reflections on the origin of the genetic code: a hypothesis. *J. Theor. Biol.* **191**, 191–196.

Dillon, L. S. 1973 The origins of the genetic code. *Bot. Rev.* **39**, 301–345.

Epstein, C. J. 1966 Role of the amino acid 'code' and of selection for conformation in the evolution of proteins. *Nature* **210**, 25–28.

Fitch, W. M. 1966 The relationship between frequencies of amino acids and ordered trinucleotides. *J. Molec. Biol.* **16**, 1–8.

Fitch, W. M. & Upper, K. 1987 The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 759–767.

Freeland, S. J. & Hurst, L. D. 1998 The genetic code is one in a million. *J. Molec. Evol.* (In the press.)

Goldberg, A. L. & Wittes, R. E. 1966 Genetic code: aspects of organisation. *Science* **153**, 420–424.

Haig, D. & Hurst, L. D. 1991 A quantitative measure of error minimization in the genetic code. *J. Molec. Evol.* **33**, 412–417.

Hartman, H. 1975 Speculations on the origin of the genetic code. *J. Molec. Evol.* **40**, 541–544.

Jukes, T. H. 1983 Evolution of the amino acid code: inferences from mitochondrial codes. *J. Molec. Evol.* **19**, 219–225.

Jukes, T. H. & Osawa, S. 1991 Recent evidence for the evolution of the genetic code. In *Evolution of life: fossils, molecules and culture* (ed. S. Osawa & T. Honjo), pp. 79–95. Tokyo: Springer.

Jukes, T. H. & Osawa, S. 1997 Further comments on codon reassignment. *J. Molec. Evol.* **45**, 1–3.

Nirenberg, M. W., Jones, O. W., Leder, P., Clark, B. F. C., Sly, W. S. & Pestka, S. 1963 On the coding of genetic information. *Cold Spring Harb. Symp. Quant. Biol.* **28**, 549–557.

Osawa, S. & Jukes, T. H. 1989 Codon reassignment (codon capture) in evolution. *J. Molec. Evol.* **28**, 271–278.

Osawa, S. & Jukes, T. H. 1995 On codon reassignment. *J. Molec. Evol.* **41**, 247–249.

Osawa, S., Jukes, T. H., Watanabe, K. & Muto, A. 1992 Recent evidence for the evolution of the genetic code. *Microbiol. Rev.* **56**, 229–264.

Pelc, S. R. & Welton, M. G. E. 1966 Stereochemical relationship between coding triplets and amino-acids. *Nature* **209**, 868–870.

Santos, M. A. S., Ueda, T., Watanabe, K. & Tuite, M. F. 1997 The non-standard genetic code of *Candida* spp.: an evolving genetic code or a novel mechanism for adaptation? *Molec. Microbiol.* **26**, 423–431.

Schultz, D. W. & Yarus, M. 1994 Transfer RNA mutation and the malleability of the genetic code. *J. Molec. Biol.* **235**, 1377–1380.

Schultz, D. W. & Yarus, M. 1996 On malleability in the genetic code. *J. Molec. Evol.* **43**, 597–601.

Sokal, R. R. & Rohlf, F. J. 1981 *Biometry*, 2nd edn. New York: W. H. Freeman.

Sonneborn, T. M. 1965 Degeneracy of the genetic code: extent, nature and genetic implications. In *Evolving genes and proteins* (ed. V. Bryson & H. J. Vogel), pp. 377–397. New York: Academic Press.

Szathmary, E. 1991 Four letters in the genetic alphabet: a frozen evolutionary optimum? *Proc. R. Soc. Lond.* B **245**, 91–99.

Szathmary, E. 1992 What is the optimum size for the genetic alphabet? *Proc. Natn. Acad. Sci. USA* **89**, 2614–2618.

Szathmary, E. 1993 Coding coenzyme handles: a hypothesis for the origin of the genetic code. *Proc. Natn. Acad. Sci. USA* **90**, 9916–9920.

Szathmary, E. & Maynard Smith, J. 1995 *The major transitions in evolution*. Oxford: W. H. Freeman.

Szathmary, E. & Zintzaras, E. 1992 A statistical test of hypotheses on the organisation and origin of the genetic code. *J. Molec. Evol.* **35**, 185–189.

Taylor, F. J. R. & Coates, D. 1989 The code within the codons. *BioSystems* **22**, 177–187.

Woese, C. R. 1965 On the evolution of the genetic code. *Proc. Natn. Acad. Sci. USA* **54**, 1546–1552.

Woese, C. R., Dugre, D. H., Dugre, S. A., Kondo, M. & Saxinger, W. C. 1966 On the fundamental nature and evolution of the genetic code. *Cold Spring Harb. Symp. Quant. Biol.* **33**, 723–736.

Wong, J. T. 1975 A co-evolution theory of the genetic code. *Proc. Natn. Acad. Sci. USA* **72**, 1909–1912.

Wong, J. T. 1976 The evolution of a universal genetic code. *Proc. Natn. Acad. Sci. USA* **73**, 2336–2340.

Wong, J. T. 1980 Role of minimisation of chemical distances between amino acids in the evolution of the genetic code. *Proc. Natn. Acad. Sci. USA* **77**, 1083–1086.

Wong, J. T. 1981 Coevolution of the genetic code and amino acid biosynthesis. *Trends Biochem. Sci.* **6**, 33–36.

Wong, J. T. 1988 Evolution of the genetic code. *Microbiol. Sci.* **5**, 174–181.

Wong, J. T. & Bronskill, P. M. 1979 Inadequacy of pre-biotic synthesis as origin of proteinous amino acids. *J. Molec. Evol.* **13**, 115–125.

Yarus, M. & Schultz, D. W. 1997 Response. *J. Molec. Evol.* **45**, 3–6.

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.